Linear Array Image Analysis
For Automated Detection of Human Papillomavirus

Matthew Wilhelm, Brian Nutter
Elec. & Comp. Engr.
Texas Tech University
brian.nutter@ttu.edu

Rodney Long, Sameer Antani
Lister Hill Nat. Cen. for Biomed. Comm.
National Library of Medicine, NIH

Abstract

Persistent infections with carcinogenic Human Papillomavirus (HPV) are a necessary cause for cervical cancer, which is the fifth most deadly cancer for women worldwide. Approximately 20 million Americans are currently infected with HPV but only a subset will develop cervical cancer. While a negative HPV test indicates a very low risk for cervical cancer, a positive test cannot discriminate between an innocuous transient infection and a prevalent cancer. Additional information such as HPV genotype and HPV viral load is thought to improve the predictive ability of which women will develop cervical cancer. The visual interpretation of hybridization-strip-based HPV genotyping results, however, is heterogeneous and poorly standardized. This has led to work toward the development of a robust automated image analysis package for HPV genotyping strips.

1. Introduction

The objective of this project is to develop an automated system for the detection of human papillomavirus (HPV) in images [3] captured from Roche Molecular Diagnostics' Linear Array (LA) HPV Genotyping Test [4]. This test provides type-specific HPV genotype results for thirty seven different types, with different risk levels for developing cervical cancer. The LA test is based on DNA amplification of a region in the HPV L1 gene followed by type-specific hybridization on strips that carry probes for 37 HPV types. Figure 1(a) below shows a set of 15 individual LA strips, which will be referred to as an LA image, where each white vertical lane contains data for an individual patient, and the dark horizontal bars inside the lanes correspond with specific types of HPV. Figure 1(b) shows the reference guide that is used to match the vertical locations of the horizontal bars in the LA image to specific types of HPV. The images captured from these hybridization strips exhibit a number of characteristics that make automated processing difficult. While these issues are minor or even unnoticeable to the naked eye, they must be resolved correctly to insure accurate image analysis. The most visible of these problems is the background level, i.e. the intensity of the white lane, in the image. This intensity level varies strongly within a given image and among multiple images. The horizontal bars indicating specific types are not of constant intensity; however, detection of both the low and high contrast bars is equally important because the effect of individual types is not yet fully understood. Another noticeable problem in the zoomed section, shown in Figure 1(c), is the presence of impulse and pattern noise, which is likely attributable to the reflectivity of the strips and to characteristics of the imaging system.

One slightly less noticeable problem is that the lanes are not always perfectly straight up and down in the photographic images; in fact, the whole image has some amount of skew, rotation, and other forms of distortion common in camera images.

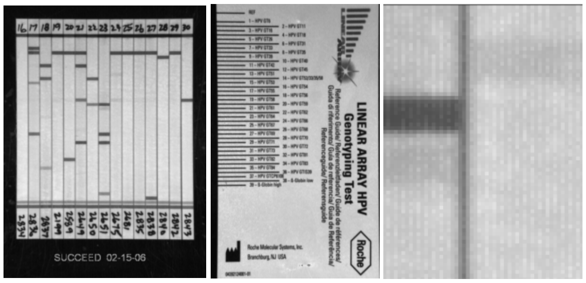


Figure 1: (a) Linear Array (LA) Genotyping hybridization strips (b) Genotyping Test Reference Guide
(c) Fragment of One Lane Cropped/Zoomed To Show Detail

The algorithms utilized in this paper are generalized, i.e. not specific to a certain test image, allowing analysis of data from various laboratories where different electrophoresis or photographic protocols may be in use. With this capability, clinical algorithms

978-1-4244-4878-4/09/$25.00 ©2009 IEEE

can be derived and, ultimately, critical patient care may be improved.

This paper begins with a discussion of goals of the project, namely robust analysis of the LA images. Section three presents the techniques that have been implemented. Section four provides a review of the results accomplished. The paper is concluded in section five with a discussion of the state of the project and planned future work.

2. Motivation

HPV genotyping is being studied extensively in various laboratories, using different assays and different evaluation protocols. Several assays are based on strip hybridization to detect HPV genotypes, similar to LA. Currently, these are evaluated visually with the goal of achieving a binary result: *type present/not present*. However, visual evaluation of genotyping strips is poorly standardized and may be very heterogeneous [3]. Furthermore, there is quantitative information (signal intensity) on these strips that is currently not used. The goal of this project is to develop an automated image analysis software package that would produce both accurate and repeatable results. Such a system would provide a tool for biomedical research and possible clinical practice. Precise genotyping information is important for natural history studies and may in the future become part of clinical management. Large randomized trials have shown that HPV testing can be efficiently used in primary cervical cancer screening. Genotyping data is important to determine which types should be included in screening assays.

To achieve uniform analysis, the algorithms are chosen and designed such that they do not make detailed assumptions about the features of the available calibration images. For example, any geometric distortion present must be corrected for each individual image. More subtle assumptions of image characteristics such as specific patterns of variation in illumination must also be avoided.

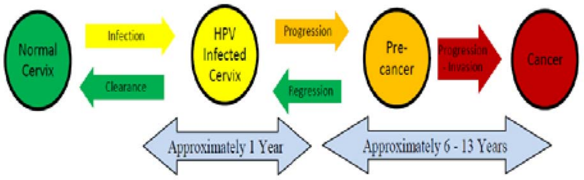


Figure 2: HPV Progression to Cervical Cancer (2)

HPV, which currently affects 20 million Americans [1], when left untreated is likely to progress to cervical cancer. Cervical Cancer is the fifth most deadly cancer in women worldwide, therefore early detection and treatment is paramount [2].

3. Implementation

The image data for this work, one of which is shown in figure 1(a), are obtained from experiments conducted by the National Cancer Institute and the University of Oklahoma and cover patients with a wide range of cervical disease, as detailed in [3]. MATLAB is used as the programming language for the automated analysis of these images, because of its inherent speed of matrix processing and its ample supply of prebuilt generic image processing algorithms. In order for the system to be useful outside of the engineering environment, it must tolerate a high degree of variability with respect to factors such as lane geometry in the image (e.g., shifts and rotation), illumination, and spatial resolution.

Several major steps must be accomplished in order to develop a solution that will robustly analyze image data acquired from a variety of labs. These steps include: removal of noise, location of patient lane dividers, transformation of single patient data into a more easily analyzed 1-D signal, analysis of the reference guide, detection of background variation while leaving HPV strands unchanged, and finally matching of patient data to reference guide location to assign HPV genotype names. Each step will be discussed thoroughly below, followed by a review of the results in the next section.

3.1 Noise removal algorithm

In order for more complex image processing techniques to work correctly, noise in the images should be reduced. Strong impulse noise if present must be removed. The median filter is a very good candidate for impulse noise removal, because it removes outliers without adversely affecting the fine details. The median filter is implemented by traversing the image pixel by pixel and setting each pixel to the median value of the 3x3 neighborhood of surrounding pixels. Figure 3 shows the results of the median filter. While the impulse noise is greatly reduced, the details are not distorted.

After the impulse noise has been removed, the software can then move to subsequent tasks such as segmenting the image in order to produce a signal for each patient that can be analyzed to detect specific types of HPV.

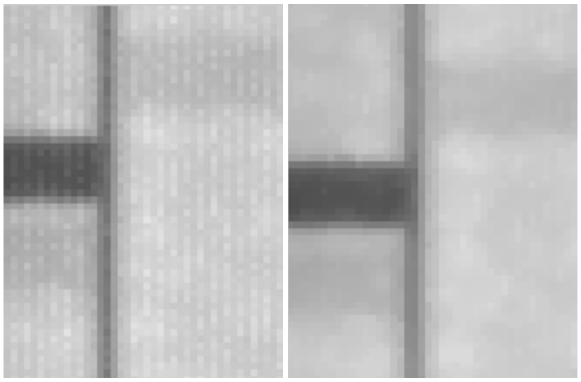


Figure 3: Impulse Noise in Right Image,
Removed in Left Image

3.2 Patient location mapping algorithm

A lane divider tracking method was developed that identifies and utilizes the locations of the tops and bottoms of the lane dividers. The ends of the lane dividers are found with a matched filter searching for a "T shape" in the image. The result of this filter is shown in Figure 4(a), where the white areas are areas that match the T filter. The local maxima in the white areas are the exact locations of the tops and bottoms of the lanes. The first 100 maxima in the resulting image are found by searching for the global maximum, storing this location, creating a 30 pixel exclusion region around it, and repeating this process 100 times; this process can be seen in figure 4(b). The exclusion region is used to avoid finding false local maxima that occur close to true local maxima. Following this stage, the rows with the highest number of maxima are clearly the location of the top and bottom rows of the region of interest.

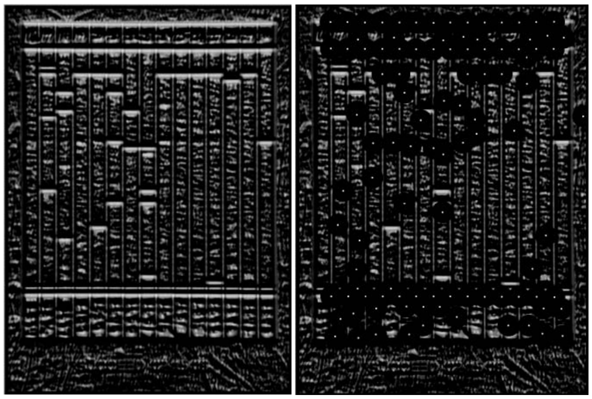
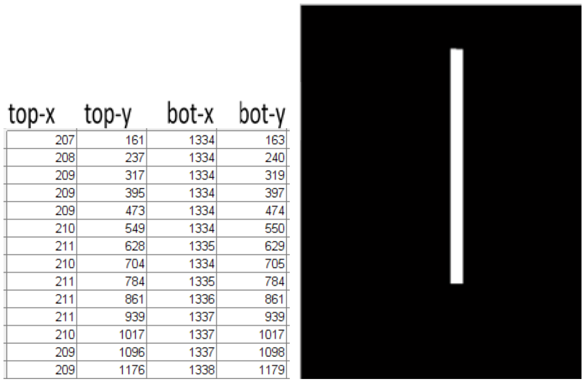


Figure 4: (a) Normalized Cross Correlation Searching for T's
(b) Locations of T's Covered by exclusion region



top-x	top-y	bot-x	bot-y
207	161	1334	163
208	237	1334	240
209	317	1334	319
209	395	1334	397
209	473	1334	474
210	549	1334	550
211	628	1335	629
210	704	1334	705
211	784	1335	784
211	861	1336	861
211	939	1337	939
210	1017	1337	1017
209	1096	1337	1098
209	1176	1338	1179

Figure 4: (c) Locations of matching top and bottom lane ends
(d) Region of interest for patient 24

The locations of the tops and bottoms of each lane divider can then be found and matched as shown in figure 4(c). The tops of the lane dividers are readily identified as existing near row 209. The bottoms of the lane dividers are near row 1335. All potential line dividers not near these rows are eliminated. The remaining lane dividers are sorted by column, and the lanes are thereby identified. This data is then used to create a custom region of interest (ROI) for each patient, which contains the image data for the HPV signal responses for that patient. Figure 4(d) shows one such ROI for a patient near the middle of the hybridization strip.

3.3 Transform patient data to 1-D signal

The patient ROI is used as a map to guide transformation of the 2-D image data into a patient-specific 1-D signal. Using this map, each patient's lane is converted into a one-dimensional signal by computing the average horizontal value of each pixel row in the ROI. Averaging the available data helps reduce the effect of remaining noise, thereby increasing the signal-to-noise ratio. These results are detailed in section 4.

3.4 Reference guide analysis

The reference guide shown in figure 1(b) is automatically analyzed to acquire information on the vertical locations of HPV types. This allows us to create a correspondence between numerical signal values from the image and the detected HPV types. Matched filtering is used once again, searching for a '—' shape to detect the horizontal lines pointing to the locations. The result produced by correlation of the mask with the reference guide is noisy because other objects in the reference guide also match the filter.

These unwanted matches are removed easily because they are much smaller than the true matches, and a clean location map is left. This map is then quantized by computing the average vertical location of each of the remaining matches. Table 1 below shows the result of the quantization. These values are later scaled to match each patient's 1-D signal to a specific type of HPV, and are used to find the background levels by looking in-between them.

Table 1: Non-scaled Reference Guide Locations

60.89	89.99	104.45	118.90	133.13	147.73	162	176.23
190.79	205	219	233.42	247.91	262	276	290
304.59	319	333	347	361	375	389	403.6
418	432	446	460	474	488	502	516.07
530.43	544.64	558.82	573	587	601	615.2	629.37

3.5 Background variation detection and resolution

The background variations are not consistent across the entire image. Our strategy is to address this issue after the data for individual patients has been separated. With the one-dimensional patient data provided from the map above, and the reference guide locations, our approach is to estimate the background level by using the fact that in-between the locations shown in table 1 there is only background. For example, the first location that is known to be background is in-between the second and third values shown in table 1, namely 89.99 and 104.45, by averaging these two values, it is known that at location 97.22 only background levels exist. The goal is to remove the background with enough specificity that even weak types of HPV will be detectable by searching for characteristic "troughs" in the data.

3.6 HPV genotype identification

With an accurate map of the reference guide locations and a de-noised patient signal, this algorithm searches for locations with both a reference guide coordinate and a patient signal showing a trough. Since this location is a known location on the reference guide, it indicates a specific type of HPV. This matching is done in two steps, simply locating possible troughs and then verifying proper shape of these troughs to remove false indications caused by noise. For example, after the initial trough location for patient 17, fourteen locations were identified as simple troughs these were types 2, 3, 6, 7, 8, 10, 12, 13, 14, 17, 22, 33, 36, and 37. While some of these are clearly correct, most of them are false indications. After checking these locations for proper depth and width, it was found that only five of these possibilities were actually types of HPV. Namely types, 2, 3, 6, 10, and 22, were identified as patient 17's diagnosis. One should notice that although type 6 is extremely week type, it is still preserved in the final diagnosis.

4. Results

Figure 5 shows patient number 24's results after being averaged to the 1-D signal. Following the graph from left to right, the beginning of the signal, right after the x=200 on the x-axis, corresponds with the first horizontal bar directly below the handwritten letters in the original image. (This bar is actually a reference bar to align the electrophoresis gel with the reference guide.) The trough at about 310 on the x-axis represents the first type of HPV detected in this patient. At about 450 and 620 on the x-axis, two other HPV troughs are visible in the 1-D plot. Patient 24's lane has been included below the plot for comparison, where it is displayed after a rotation of 90 degrees. While the second and third HPV types are clearly visible in the 1-D plot, these types are not so obvious when looking at the original image. An important observation to note here is that while the trough at 310 is much deeper than the troughs at 450 and 620, the width of all three troughs is very similar, as are the widths of the corresponding horizontal bars in the original image. This width similarity will be utilized to make detection of the trough locations relatively tractable and reasonably robust to noise.

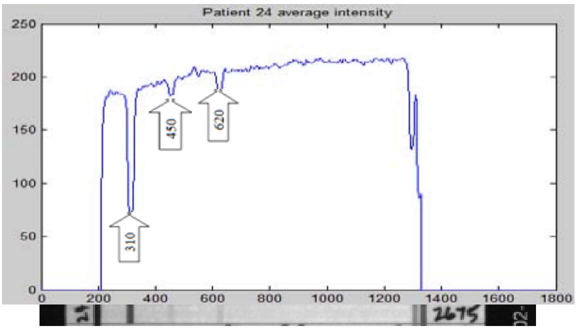


Figure 5: Patient #24 Detailed Results

Figure 6 shows detailed results for patient 17. This case illustrates two HPV strains that are located at adjacent positions in the gel, so that identification of each strain appears possible, if nontrivial. These two strains occur at about locations 280 and 310.

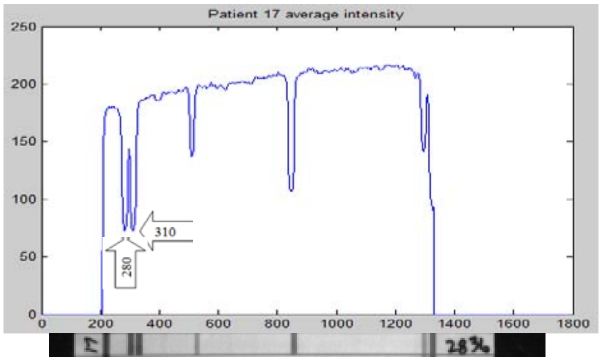


Figure 6: Patient #17 Detailed Results

Figure 7 shows detailed results for patient 16. This case is interesting because no HPV strains can be identified, and background and noise can be assessed.

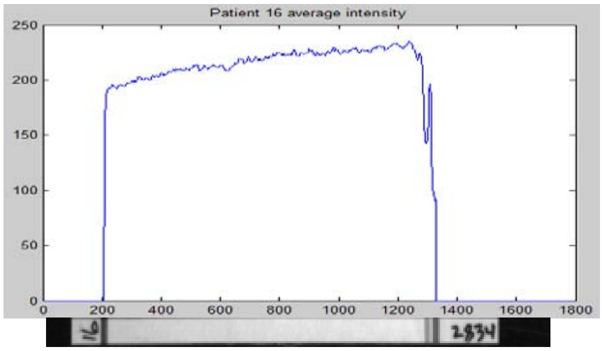


Figure 7: Patient #16 Detailed Results

5. Conclusion

The algorithms mentioned above have been implemented and tested on a variety of patient data, and have been show to be robust and repeatable. The system developed, provides genotype specific classification of the HPV types present in each patient.

Future work includes making a user friendly clinical quality software package. Other future work that would be helpful to researchers and medical experts would be the development of a medical database to store the resulting data and to track the changes in individuals' records over several years. This database could be utilized to develop clinical algorithms that would suggest actions based on previous patient outcomes.

Acknowledgments

The authors gratefully acknowledge the contribution of HPV data from the Study to Understand Cervical Cancer Early Endpoints and Determinants (SUCCEED), provided by Terence Dunn and R. Andy Allen, of the University of Oklahoma Health Sciences Center. We also thank Dr. Nicolas Wentzensen and Sophia Wang of the National Cancer Institute for this data, and for their review of, and contributions to, this paper. This research was supported by the Intramural Research Program of the U.S. National Institutes of Health, National Library of Medicine, and Lister Hill National Center for Biomedical Communications, Contract No. HHSN276200800411P.

6. References

1. Human Papilomavirus. *CDC - STD Facts.* [Online] February 14, 2009. [Cited: February 14, 2009.] http://www.cdc.gov/std/HPV/STDFact-HPV.htm.

2. Cancer Fact Sheet. *World Health Organization.* [Online] February 2009. [Cited: February 14, 2009.] http://www.who.int/mediacentre/factsheets/fs297/en/index.html.

3. *Evaluation of Linear Array Human Papillomavirus Genotyping Using Automatic Optical Imaging Software.* Jeronimo, J., et al. 8, s.l. : American Society for Microbiology, 2008, Vol. 46.

4. Diagnostics Products - HPV. *Roche Molecular Dignostics Global.* [Online] February 2009. [Cited:February 12, 2009.] http://molecular.roche.com/diagnostics/hpv_ctng/hpv_test_2.html.